BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images may include (but are not limited to):

- BLACK BORDERS
- TEXT CUT OFF AT TOP, BOTTOM OR SIDES
- FADED TEXT
- ILLEGIBLE TEXT
- SKEWED/SLANTED IMAGES
- COLORED PHOTOS
- BLACK OR VERY BLACK AND WHITE DARK PHOTOS
- GRAY SCALE DOCUMENTS

IMAGES ARE BEST AVAILABLE COPY.

As rescanning documents *will not* correct images, please do not report the images to the Image Problem Mailbox.

(19) JAPANESE PATENT OFFICE

PATENT ABSTRACTS OF JAPAN

(11) Publication number: **10105572 A**

(43) Date of publication of application: **24.04.98**

(51) Int. Cl **G06F 17/30**
G06F 17/21

(21) Application number: **08262047**

(22) Date of filing: **02.10.96**

(71) Applicant: **NEC CORP**

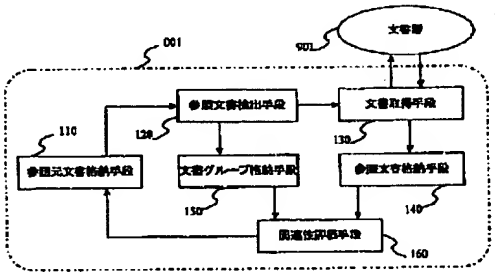
(72) Inventor: **YAMAGUCHI TOMOHARU**

(54) **DEVICE AND METHOD FOR GROUPING DOCUMENTS**

(57) Abstract:

PROBLEM TO BE SOLVED: To determine a collection object range limiting the range to follow the reference relation so as to collect documents having a semantically deep relationship when collecting documents based on the reference relation between documents.

SOLUTION: A reference document detecting means 120 extracts a document stored in a reference source document storage means 110, detects the reference relation with the other document out of this document, and stores the completely detected source document in a document group storage means 150. A document possessing means 130 possesses the document, which is detected by the reference document detecting means 120, having the reference relation from a document group and stores this document in a reference document storage means, and a relationship evaluating means evaluates the relationship between the reference document stored in the reference document storage means 140 and the document group stored in the document group storage means 150. When there is any relationship, the reference document is added to the reference source document storage means 140 as a new reference source document, and operation is repeated from processing at the reference document detecting means 120. When there is no document stored in the reference source document storage means 120, it is determined the documents stored in the document storage means 130 belong to one group.



(19)日本国特許庁（ＪＰ）　(12)公開特許公報（A）　(11)特許出願公開番号
特開平10－105572
(43)公開日　平成10年(1998)４月24日

(51)Int.Cl.[6]	識別記号	FI		
G06F	17/30	G06F	15/401	310D
	17/21		15/20	570N
			15/40	370A

審査請求　有　請求項の数12　OL　（全９頁）

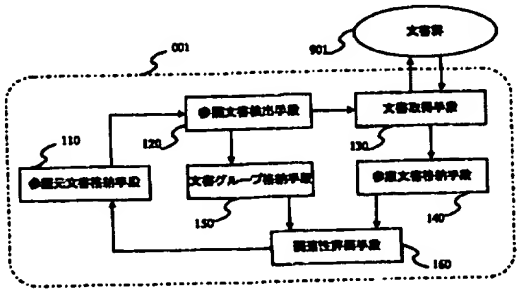
(21)出願番号　特願平8－262047

(22)出願日　平成８年(1996)10月２日

(71)出願人　000004237
日本電気株式会社
東京都港区芝五丁目７番１号
(72)発明者　山口　智治
東京都港区芝五丁目７番１号　日本電気株式会社内
(74)代理人　弁理士　京本　直樹　（外２名）

(54)【発明の名称】　文書グループ化装置および文書グループ化方法

(57)【要約】
【課題】　文書の参照関係に基づいた文書の収集において、意味的に関連性の深い文書を収集できるように参照関係を辿る範囲を限定した収集対象範囲を決定する。
【解決手段】　参照文書検出手段は、参照元文書格納手段に格納された文書を取り出して該文書中から他の文書への参照関係を検出し、検出し終えた参照元文書を文書グループ格納手段に格納する。文書取得手段は、参照文書検出手段により検出された参照関係のある文書を文書群から取得して参照文書格納手段に格納し、関連性評価手段は、参照文書格納手段に格納された参照文書と文書グループ格納手段に格納された文書群の関連性を評価し、関連がある場合には参照文書を参照元文書格納手段に新たな参照元文書として追加し、参照文書検出手段の処理から繰り返す。参照元文書格納手段に格納されている文書が無くなれば、文書格納手段に格納されている文書を一つのグループに属するものと決定する。



1

【特許請求の範囲】

【請求項１】文書間の参照関係情報を持ち、分散して存在する文書群について、分散された文書を収集する際に、任意の特定の文書を起点として、該文書から参照関係を辿って到達可能な文書群のうち、該文書に関連の深い文書のみを効率よく収集するために、参照関係に基づいて参照される文書と、収集済みの文書群との関連性を判定することにより、参照される文書を収集対象とするか否かを決定することで、参照関係を辿る範囲を限定することを特徴とする文書グループ化装置。

【請求項２】分散して存在する文書群の中から、ある文書に関連性が有る文書を収集して、グループ化する文書グループ化装置において、

ある文書（参照元文書）を起点として、前記参照元文書中に存在する他の文書（参照文書）の参照関係情報を取り出して、前記参照関係情報に基づいた参照文書を収集し、

前記参照元文書と前記参照文書の関連性を判定し、前記参照文書が前記参照元文書と関連性がある場合には、前記参照文書を参照元文書に追加し、さらに追加された参照元文書から参照可能な文書の関連性を判定することによって、ある文書に関連性のある文書を収集して、グループ化することを特徴とする文書グループ化装置。

【請求項３】分散して存在する複数のハイパーテキスト文書の中から、あるハイパーテキスト文書に関連性のある他のハイパーテキスト文書を収集して、関連性があるハイパーテキスト文書同士をグループ化する文書グループ化装置において、

前記ハイパーテキスト文書中のリンクを順次辿って関連文書を収集する際に、リンク先の文書と収集済みの文書との関連性を判定することにより、リンク先の文書を収集対象とするか否かを決定し、収集対象としない場合には、そのリンク先の文書からのリンクも辿らないことにすることで、あらかじめ特定の収集条件を設定することなくリンクを辿る範囲を限定することを特徴とする文書グループ化装置。

【請求項４】ＨＴＭＬ形式で記述されたハイパーテキスト文書が、ネットワークを介して複数の計算機内に存在し、ある特定のハイパーテキスト文書に関連性が有る文書を前記計算機から収集して、関連性のある文書同士をグループ化する文書グループ化装置において、

文書収集の起点となるハイパーテキスト文書（参照元文書）から、他のハイパーテキスト文書の参照を示すＵＲＬを検出し、ネットワークを介して前記ＵＲＬに該当するハイパーテキスト文書（参照文書）を収集し、前記参照元文書と収集された前記参照文書の関連性が有るか無いかを判断し、前記参照文書が関連性が有りと判断されたものは、前記参照文書を参照元文書として追加し、さらに追加された参照元文書内のＵＲＬを検出して、他の参照文書を収集する動作を繰り返すことにより、関連性のあるハイパーテキスト文書をグループ化することを特徴とする文書グループ化装置。

2

【請求項５】請求項４に記載された文書グループ化装置において、

前記参照元文書と収集された前記参照文書の関連性が有るか無いかを判断する時に、前記参照元文書と前記参照文書に含まれるキーワードを抽出し、前記キーワードの一致度により、関連性を判断することを特徴とする文書グループ化装置。

【請求項６】文書収集の起点となる文書を格納する参照元文書格納手段と、

前記参照元文書格納手段に格納された文書を順次取り出して該文書中から他の文書への参照関係を記述した箇所を検出する参照文書検出手段と、

前記参照文書検出手段により検出された参照関係により、前記参照関係に対応する文書を文書群から取得する文書取得手段と、

前記文書取得手段により取得された文書を格納しておく参照文書格納手段と、

前記参照文書検出手段により文書中の参照関係を検出し終えた参照元文書を参照元文書格納手段から移して格納しておく文書グループ格納手段と、

前記参照文書格納手段に格納された参照文書と文書グループ格納手段に格納された文書群の関連性を評価し、関連がある場合には参照文書を参照元文書格納手段に新たな参照元文書として追加する関連性評価手段とを含んで構成され、

文書の参照関係に基づいた文書の収集において、到達可能な全文書を探索することなく意味的に関連性の深い文書を収集できるように、参照関係を辿る範囲を限定した収集対象範囲を決定をする文書グループ化装置。

【請求項７】分散して存在する文書群の中から、ある文書に関連性が有る文書を収集して、グループ化する文書グループ化方法において、

参照元文書格納手段に格納している文書（参照元文書）を起点として、前記参照元文書中に存在する他の文書（参照文書）の参照関係情報を取り出す第１のステップと、

前記参照元文書を文書グループとして文書グループ格納手段に格納する第２のステップと、

前記第１のステップにより取り出された参照関係情報により、前記文書群から参照文書を取得する第３のステップと、

前記第２のステップで格納された文書グループの参照元文書と、前記第３にステップにより取得された参照文書との関連性の有り無しを判断する第４のステップと、

前記第４のステップにより関連性が有りと判断された参照文書を、参照元文書として前記参照元文書格納手段に追加する第５のステップと、

前記参照元文書格納手段に、参照関係情報が取り出され

ていない参照元文書が有るか無いかを判断し、参照元文書が有る場合には前記第１のステップに戻り一連の動作を繰り返し、参照元文書が無い場合には、得られた文書グループによりグループ化を決定する第６のステップと、
を備えることを特徴とする文書グループ化方法。
【請求項８】分散して存在する文書群の中から、ある文書に関連性が有る文書を収集して、グループ化する文書グループ化プログラムを記録した記録媒体において、
第１の記憶手段内に格納している文書（参照元文書）を起点として、前記参照元文書中に存在する他の文書（参照文書）の参照関係情報を取り出す第１のステップと、
前記参照元文書を文書グループとして第２の記憶手段に格納する第２のステップと、
前記第１のステップにより取り出された参照関係情報により、前記文書群から参照文書を取得する第３のステップと、
前記第２のステップで格納された文書グループの参照元文書と、前記第３にステップにより取得された参照文書との関連性の有り無しを判断する第４のステップと、
前記第４のステップにより関連性が有りと判断された参照文書を、参照元文書として前記第１の記憶手段に追加する第５のステップと、
前記第１の記憶手段に、参照関係情報が取り出されていない参照元文書が有るか無いかを判断し、参照元文書が有る場合には前記第１のステップに戻り一連の動作を繰り返し、参照元文書が無い場合には、得られた文書グループによりグループ化を決定する第６のステップと、
を少なくとも備えるプログラムを格納したことを特徴とする文書グループ化プログラムを記録した記録媒体。
【請求項９】文書収集の起点となる文書を格納する参照元文書格納手段と、
前記参照元文書格納手段に格納された文書を順次取り出して該文書中から他の文書への参照関係の説明を記述した箇所を検出する参照文書検出手段と、
前記参照文書検出手段により文書中の参照関係の説明を検出し終えた参照元文書を前記参照元文書格納手段から移して格納しておく文書グループ格納手段と、
前記参照文書検出手段により検出された参照関係の説明に該当する文書中の記述と文書グループ格納手段に格納された参照元文書の本文の記述とから、前記参照元文書と参照関係のある文書との関連性を評価する関連性評価手段と、
前記関連性評価手段により関連があると評価された文書のみを文書群から取得する文書取得手段と、
前記文書取得手段により取得された文書を格納しておく参照文書格納手段と、から構成され、
前記関連性評価手段は、前記参照文書格納手段が格納した文書を参照元文書として、前記参照元文書格納手段に格納する処理を更に行うことを特徴とする文書グループ化装置。
【請求項１０】ＨＴＭＬ形式で記述されたハイパーテキスト文書が、ネットワークを介して複数の計算機内に存在し、ある特定のハイパーテキスト文書に関連性のある文書を前記計算機から収集して、関連性のある文書同士をグループ化する文書グループ化装置において、
文書収集の起点となるハイパーテキスト文書（参照元文書）から、他のハイパーテキスト文書（参照文書）を示すＵＲＬを検出し、
前記ＵＲＬを説明している文字列のキーワードと、前記参照元文書の本文中のキーワードの一致度を算出することによって、前記参照元文書と前記参照文書の関連性が有るか無いかを判断し、関連性が有ると判断されたＵＲＬで示される参照文章を前記ネットワークを介して該計算機から得ることによって関連性のある文書を収集し、さらに、得られた参照文章を参照元文書として他の参照文書を収集する動作を繰り返すことによって、前記参照元文書に関連性が有る文書同士をグループ化することを特徴とする文書グループ化装置。
【請求項１１】分散して存在する文書群の中から、ある文書に関連性が有る文書を収集して、グループ化する文書グループ化方法において、
参照元文書格納手段に格納している文書（参照元文書）を起点として、前記参照元文書中に存在する他の文書（参照文書）の参照関係情報の説明を取り出す第１のステップと、
前記参照元文書を文書グループとして文書グループ格納手段に格納する第２のステップと、
前記第１のステップにより取り出された参照関係情報の説明の内容と、前記参照元文書の本文の内容との関連性の有り無しを判断する第３のステップと、
前記第３のステップにより関連性が有りと判断された参照文書を、前記文書群から取得する第４のステップと、
前記第４のステップにより取得された参照文書を、参照元文書として前記参照元文書格納手段に追加する第５のステップと、
前記参照元文書格納手段に、参照関係情報が取り出されていない参照元文書が有るか無いかを判断し、参照元文書が有る場合には前記第１のステップに戻り一連の動作を繰り返し、前記参照元文書が無い場合には、得られた文書グループによりグループ化を決定する第６のステップと、
を備えることを特徴とする文書グループ化方法。
【請求項１２】分散して存在する文書群の中から、ある文書に関連性が有る文書を収集して、グループ化する文書グループ化プログラムを記録した記録媒体において、
第１の記憶手段内に格納している文書（参照元文書）を起点として、前記参照元文書中に存在する他の文書（参照文書）の参照関係情報を取り出す第１のステップと、
前記参照元文書を文書グループとして第２の記憶手段に

5

格納する第２のステップと、
前記第１のステップにより取り出された参照関係情報の説明の内容と、前記参照元文書の本文の内容との関連性の有り無しを判断する第３のステップと、
前記第３のステップにより関連性が有りと判断された参照文書を、前記文書群から取得する第４のステップと、
前記第４のステップにより取得された参照文書を、参照元文書として前記第１の記憶手段に追加する第５のステップと、
前記第１の記憶手段に、参照関係情報が取り出されていない参照元文書が有るか無いかを判断し、参照元文書が有る場合には前記第１のステップに戻り一連の動作を繰り返し、前記参照元文書が無い場合には、得られた文書グループによりグループ化を決定する第６のステップと、
を少なくとも備えるプログラムを格納したことを特徴とする文書グループ化プログラムを記録した記録媒体。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は電子化された文書の収集装置に関し、特に分散された複数の文書を関連性のあるものとないものに区別して収集範囲を決定し、関連性のある文書を収集してグループ化を行う文書グループ化装置および文書グループ化方法、さらに、文書グループ化を行うプログラムを記録した記録媒体に関する。

【０００２】

【従来の技術】従来、分散された文書間の参照関係情報をもつ文書群の代表的なものとしてハイパーテキストが知られている。ハイパーテキストは複数の文書をリンクによって関連付けることが可能である。

【０００３】この代表的な従来技術の例として、特開平４－３２１１４４号公報に記載の「ハイパーテキストのブラウジング処理装置」が知られている。この従来技術によれば、文書間のリンク付け関係を一覧することができる。

【０００４】また、特開平５－１２８１５７号公報に記載の「文書検索装置」によれば、リンクを辿って到達可能な特定の範囲から、指定したキーワードにマッチする内容をもつ文書を選択的に検索することができる。

【０００５】さらに、（株）エーアイソフトの「Ｗｅｂ Ｗｈａｃｋｅｒ」（米国ＦｏｒｅＦｒｏｎｔ Ｇｒｏｕｐ，Ｉｎｃ．の商標）、株式会社ビー・ユー・ジーの「ＰｅｒＭａｎ Ｓｕｒｆｅｒ波乗野郎」（株式会社ビー・ユー・ジーの商標）などに代表されるのオートパイロットあるいはダウンローダと呼ばれるソフトウェアによれば、大規模なハイパーテキストデータベースであるワールドワイドウェッブ（Ｗｏｒｌｄ Ｗｉｄｅ Ｗｅｂ：WWW）上の指定された文書からリンクした文書を指定された数だけ、自動収集することができる。

【０００６】

6

【発明が解決しようとする課題】前述の「ハイパーテキストのブラウジング処理装置」では文書間の意味的な関連性は表現されず、ツリー構造状に表示された文書群のどこからどこまでが、意味的に近接した関係にあるのかの判断は、人間が各文書の内容を見て判断するしかない。ワールドワイドウェッブのように大規模なハイパーテキストでは、この判断を人手でおこなうのは現実的ではない。

【０００７】また、「文書検索装置」を用いると、リンクを辿って到達可能な範囲の文書をすべて一度収集する必要があり、到達可能な範囲が膨大である場合にも、全ての文書の内容を参照する必要があり、莫大な処理時間を要するという問題を生じる。また、通信路の細いネットワーク上に分散された文書を参照する場合には、通信時間などのオーバーヘッドが大きくなるという問題を生じる。キーワード指定がリンクの作成意図と一致しない場合には、キーワードにマッチしない文書を介して間接的にリンクされているキーワードにマッチする文書間でリンクが失われてしまうという問題がある。

【０００８】オートパイロットやダウンローダなどのソフトウェアでは、辿るリンクの数や、物理的に文書が格納されているマシンによって文書の収集範囲を限定しており、文書の内容による意味的な関連性は考慮されておらず、内容的にあまり関連のない文書も収集してしまうという問題や、他のマシンに格納されている関連の深い文書が収集されないという問題がある。

【０００９】そこで、ワールドワイドウェッブのように大規模なハイパーテキストに関しても、リンクの作成意図と文書の内容に沿って関連性の深い文書に限定して収集する方法が必要である。

【００１０】本発明の目的は、文書の参照関係に基づいた文書の収集において、到達可能な全文書を探索することなく意味的に関連性の深い文書を収集できるように、参照関係を辿る範囲を限定した収集対象範囲を決定をする文書グループ化装置および文書グループ化方法と文書のグループ化が可能なプログラムを記録した記録媒体を提供することにある。

【００１１】

【課題を解決するための手段】本発明の第１の発明の文書グループ化装置は、文書収集の起点となる文書を格納する参照元文書格納手段と、参照元文書格納手段に格納された文書を順次取り出して該文書中から他の文書への参照関係を記述した箇所を検出する参照文書検出手段と、参照文書検出手段により検出された参照関係により、前記参照関係に対応する文書を文書群から取得する文書取得手段と、文書取得手段により取得された文書を格納しておく参照文書格納手段と、参照文書検出手段により文書中の参照関係を検出し終えた参照元文書を参照元文書格納手段から移して格納しておく文書グループ格納手段と、参照文書格納手段に格納された参照文書と文

書グループ格納手段に格納された文書の関連性を評価し、関連がある場合には参照文書を参照元文書格納手段に新たな参照元文書として追加する関連性評価手段とを含んで構成される。

【００１２】また、本発明の第２の発明の文書グループ化装置は、文書収集の起点となる文書を格納する参照元文書格納手段と、前記参照元文書格納手段に格納された文書を順次取り出して該文書中から他の文書への参照関係の説明を記述した箇所を検出する参照文書検出手段と、前記参照文書検出手段により文書中の参照関係の説明を検出し終えた参照元文書を前記参照元文書格納手段から移して格納しておく文書グループ格納手段と、前記参照文書検出手段により検出された参照関係の説明に該当する文書中の記述と文書グループ格納手段に格納された参照元文書の本文の記述とから、前記参照元文書と参照関係のある文書との関連性を評価する関連性評価手段と、前記関連性評価手段により関連があると評価された文書のみを文書群から取得する文書取得手段と、前記文書取得手段により取得された文書を格納しておく参照文書格納手段と、から構成され、前記関連性評価手段は、前記参照文書格納手段が格納した文書を参照元文書として、前記参照元文書格納手段に格納する処理を更に行う。

【００１３】第１の発明によれば、文書の参照関係に基づいた文書の収集において、到達可能な全文書を探索することなく意味的に関連性の深い文書を収集できるように、参照関係を辿る範囲を限定した収集対象範囲を決定を行うことが可能である。

【００１４】また、第２の発明によれば、文書の参照関係の説明によって、文書作成者の意図と文書の内容に沿って関連性の深い文書に限定した収集が可能である。

【００１５】

【発明の実施の形態】次に図１から図６を参照して本発明の実施の形態について説明する。

【００１６】図１は本発明の第１の発明である請求項１～請求項８に記載した本発明の実施の形態の一構成例を示すブロック図である。

【００１７】かかる発明の実施の形態における文書グループ化装置（００１）は、文書収集の起点となる文書を格納する参照元文書格納手段（１１０）と、参照元文書格納手段（１１０）に格納された文書を順次取り出して該文書中から他の文書への参照関係を検出する参照文書検出手段（１２０）と、参照文書検出手段（１２０）により検出された参照関係のある文書を文書群（９０１）から取得する文書取得手段（１３０）と、文書取得手段（１３０）により取得された文書を格納しておく参照文書格納手段（１４０）と、参照文書検出手段（１２０）により文書中の参照関係を検出し終えた参照元文書を参照元文書格納手段（１１０）から移して格納しておく文書グループ格納手段（１５０）と、参照文書格納手段（１４０）に格納された参照文書と文書グループ格納手段（１５０）に格納された文書群の関連性を評価し、関連がある場合には参照文書を参照元文書格納手段（１１０）に新たな参照元文書として追加する関連性評価手段（１６０）とを含んで構成される。

【００１８】また、文書グループ格納手段（１５０）格納されるのは、文書を特定できる情報のみでもよい。

【００１９】図２は請求項１から請求項８に記載した本発明の処理の流れの一実施の形態を示すフロー図である。

【００２０】参照文書検出手段（１２０）は、参照元文書格納手段（１１０）に格納された文書を順次取り出して該文書中から他の文書への参照関係を検出し（ステップＳ１０）、文書中の参照関係を検出し終えた参照元文書を参照元文書格納手段（１１０）から文書グループ格納手段（１５０）へ移して格納し（ステップＳ２０）、文書取得手段（１３０）は、参照文書検出手段（１２０）により検出された参照関係のある文書を文書群（９０１）から取得（ステップＳ３０）して参照文書格納手段（１４０）に格納し、関連性評価手段（１６０）は、参照文書格納手段（１４０）に格納された参照文書と文書グループ格納手段（１５０）に格納された文書群の関連性を評価し（ステップＳ４０）、関連がある場合には参照文書を参照元文書格納手段（１１０）に新たな参照元文書として追加（ステップＳ４０）し、参照元文書格納手段（１１０）にまだ文書が格納されているかチェックし（ステップＳ６０）、格納されている場合にはステップＳ１０から繰り返す。参照元文書格納手段（１１０）格納されている文書が無くなれば、文書グループ格納手段（１５０）に格納されている文書を一つのグループに属するものと決定する（ステップＳ７０）。

【００２１】

【実施例】以下、図面を参照して本発明の文書グループ化装置のさらに詳しい実施例について説明する。

【００２２】前述した様に、図１は、本発明の文書グループ化装置の一実施の形態の構成例を示すブロック図である。

【００２３】また、本実施例においては、文書群（９０１）として、ＨＴＭＬ形式で記述されているハイパーテキストであり、ワールドワイドウェッブのページとしてインターネットに接続された計算機上に分散して存在しているものとして説明する。

【００２４】各文書は、通信プロトコルとホスト名およびパス名を含むＵＲＬ（Ｕｎｉｆｏｒｍ　Ｒｅｓｏｕｒｃｅ　Ｌｏｃａｔｏｒ）と呼ばれる記述法により特定できる。文書取得手段（１３０）は、例えばＵＲＬに指定されたプロトコルによりインターネットに接続されている指定されたホスト計算機から指定されたパス名に該当する文書を取得する。

【００２５】例えば、参照元文書格納手段（１１０）に

格納された文書収集の起点となる文書が図３に示す文書（５０１）であるとする。ＨＴＭＬ形式の文書では参照関係情報は、”＜”と”＞”に囲まれたタグと呼ばれる部分のうち、”＜ａ”　で始まり次の”＞”までの間にある”ｈｒｅｆ＝”に続いてＵＲＬを記述し、他の文書への参照を示す箇所である。

【００２６】参照文書検出手段（１２０）は、参照元文書中から他の文書への参照を示すＵＲＬ（参照関係情報）を検出し（ステップＳ１０）、文書取得手段（１３０）により、そのＵＲＬに該当する文書を文書群（９０１）から取得して、参照文書格納手段（１４０）に格納する（ステップＳ３０）。図３の文書（５０１）からは参照文書として、ｈｔｔｐ：／／ｗｗｗ．ｆｉｓｈｅｒｍａｎ．ｃｏｍ／ｍａｒｕ．ｈｔｍｌ」および「ｈｔｔｐ：／／ｗｗｗ．ｓｈｏｐｓ．ｃｏｍ／ｆｉｓｈｉｎｇ．ｈｔｍｌ」の２つのＵＲＬが検出される。例えば、これらＵＲＬに該当する文書がそれぞれ、図４の文書（５０２）、図５の文書（５０３）に示す文書であるとする。文書中のＵＲＬを検出し終えると文書（５０１）は、文書グループ格納手段（１５０）へ移される（ステップＳ２０）。

【００２７】関連性評価手段（１６０）は例えば、文書グループ格納手段（１５０）に格納された文書（５０１）と参照文書格納手段（１４０）に格納された文書（５０２）と文書（５０３）の各文書からタグと不要語を除いてキーワードを抽出し、文書（５０１）に含まれるキーワードが文書（５０２）と文書（５０３）のそれぞれに含まれる度合いを計算して、文書（５０２）と文書（５０３）のそれぞれが文書（５０１）に対する関連性を評価する。

【００２８】本例においては、文書（５０１）の本文中のキーワードが「ＦＩＳＨＩＮＧ、釣り、フライフィッシング」であり、文書（５０２）のキーワードが「釣り、フライフィッシング、渓流釣り」であり、文書（５０３）のキーワードが「釣り、ルアー、ショップ」であるとし、関連性を参照元の文書中のキーワード全体に対する参照文書中に含まれる参照元のキーワードの数の比とし、例えば、６０％を関連性の有無を判定する基準とすれば、文書（５０２）の文書（５０１）に対する関連性は約６７％、文書（５０３）の文書（５０１）に対する関連性は約３３％となり、文書（５０２）は関連性有り、文書（５０３）は関連性無しと判定する（ステップＳ４０）。

【００２９】関連性無しと判定された文書（５０３）は、この時点で破棄され、文書（５０３）からさらに参照される文書があったとしても、それらについては取得しない。

【００３０】関連性有りと判定された文書（５０２）は、参照元文書格納手段（１１０）に新たな参照元文書として追加し（ステップＳ６０）、以下、文書（５０１）のときと同様に文書（５０２）を参照元文書として上記の過程を適用し、文書（５０２）からＵＲＬを検出し（ステップＳ１０）、さらに参照される文書を取得する（ステップＳ３０）。

【００３１】文書（５０２）からＵＲＬの検出を終え、文書グループ格納手段（１５０）に格納される（ステップＳ２０）と、文書グループ格納手段（１５０）には文書（５０１）と文書（５０２）の２つの文書が格納されている。関連性評価手段（１６０）は例えば、これら文書のキーワードの和集合を参照元のキーワード群として、文書（５０２）から検出されたＵＲＬが示す文書の関連性を評価する（ステップＳ４０）。例えばここで、検出されたＵＲＬが示す文書がいずれも関連性無しと判定され、参照元文書格納手段に参照元文書がなければ（ステップＳ６０）、処理は終了し（ステップＳ７０）、この時点で文書グループ格納手段（１５０）に格納されている文書（５０１）と文書（５０２）が、ひとつのグループをなす。

【００３２】文書グループ格納手段（１５０）の容量を節約したい場合は、文書を特定するＵＲＬとキーワード群のみを文書グループ格納手段（１５０）に格納してもよい。

【００３３】また、関連性評価手段（１６０）については、このキーワードマッチングによる実施例はあくまで一例であって、本発明は、この実施例だけに限定されるものではない。例えば、シソーラスなどを用いてキーワード間の距離を計算し、参照される文書のキーワード群の間の距離の総和や平均を用い、距離の大きさを判定基準に用いることもできる。

【００３４】次に、本発明の第２の発明である請求項９～１２に記載した発明を図面を参照して説明する。

【００３５】図６は、第２の発明の実施の形態の一構成例を示すブロック図である。本実施例においては、対象文書としてＨＴＭＬ形式のハイパーテキスト文書を扱う場合の実施例について説明する。また、先に説明した第１の発明と機能が重なる箇所については、説明を省略する。

【００３６】第２の発明の関連性評価手段（１６０）は、参照文書検出手段（１２０）により検出されるＵＲＬを説明している参照元文書中の文字列と、文書グループ格納手段（１５０）中の文書からその文字列を除いた部分との関連性を判定することで、参照される文書の関連性を推定し、関連性が有ると推定された文書についてのみ文書取得手段（１３０）により、そのＵＲＬに該当する文書を文書群（９０１）から取得して、参照文書格納手段（１４０）に格納する。

【００３７】例えば、図３の文書（５０１）では、ＵＲＬが記述されているタグ“＜ａ…＞”と対応するタグ“＜／ａ＞”との間の文字列をＵＲＬに対する説明となる文字列とみなし、ＵＲＬ「ｈｔｔｐ：／／ｗｗｗ．ｆ

11

ｉｓｈｅｒｍａｎ．ｃｏｍ／ｍａｒｕ．ｈｔｍｌ」に対して「丸山さんの釣り情報（フライフィッシングの話題もあり）」という文字列が、この参照文書（ＵＲＬ）を説明する文書になり、さらにＵＲＬ「ｈｔｔｐ：／／ｗｗｗ．ｓｈｏｐｓ．ｃｏｍ／ｆｉｓｈｉｎｇ．ｈｔｍｌ」に対して「その他の釣り情報」という文字列がこの参照文書（ＵＲＬ）を説明する文書となる。それぞれの文字列のキーワードは「丸山、釣り、フライフィッシング」、「その他、釣り」となる。

【００３８】ここで、前述した様な、参照元文書と参照文書間と同様な関連性判定を行えば、ＵＲＬ「ｈｔｔｐ：／／ｗｗｗ．ｆｉｓｈｅｒｍａｎ．ｃｏｍ／ｍａｒｕ．ｈｔｍｌ」の説明のキーワード「丸山、釣り、フライフィッシング」が、文書（５０１）の本文中に含まれている率（関連性）は約６７％となり、ＵＲＬ「ｈｔｔｐ：／／ｗｗｗ．ｓｈｏｐｓ．ｃｏｍ／ｆｉｓｈｉｎｇ．ｈｔｍｌ」の説明のキーワード「その他、釣り」が文書（５０１）の本文中に含まれている率（関連性）は５０％となる。ここで、６０％を関連性の有り無しを判定する基準とすれば、ＵＲＬ「ｈｔｔｐ：／／ｗｗｗ．ｆｉｓｈｅｒｍａｎ．ｃｏｍ／ｍａｒｕ．ｈｔｍｌ」は関連性有り、ＵＲＬ「ｈｔｔｐ：／／ｗｗｗ．ｓｈｏｐｓ．ｃｏｍ／ｆｉｓｈｉｎｇ．ｈｔｍｌ」は関連性無しと判定する。

【００３９】関連性無しと判定されたＵＲＬ「ｈｔｔｐ：／／ｗｗｗ．ｓｈｏｐｓ．ｃｏｍ／ｆｉｓｈｉｎｇ．ｈｔｍｌ」の文書（５０３）の取得はおこなわず、ＵＲＬ「ｈｔｔｐ：／／ｗｗｗ．ｆｉｓｈｅｒｍａｎ．ｃｏｍ／ｍａｒｕ．ｈｔｍｌ」の文書（５０２）のみを文書取得手段（１３０）により取得し、参照文書格納手段（１４０）に格納する。その後は前述した実施例の説明と同様に文書間の関連性判定をおこなって処理を継続する。あるいはＵＲＬの説明による関連性の推定を信頼して、文書間の関連性判定を省略することもできる。

12

【００４０】また、本発明においては、以上の述べたような構成をコンピュータプログラムによって作成し、フロッピーディスクやＣＤ－ＲＯＭに代表される記録媒体によって記録してもよい。

【００４１】

【発明の効果】本発明によれば、文書の参照関係に基づいた文書の収集において、到達可能な全文書を探索することなく意味的に関連性の深い文書を収集できるように、参照関係を辿る範囲を限定した収集対象範囲を決定をする文書グループ化装置を提供でき、ワールドワイドウェッブのように大規模なハイパーテキストに関しても、リンクの作成意図と文書の内容に沿って関連性の深い文書に限定した収集が可能になる。

【図面の簡単な説明】

【図１】本発明の文書グループ化装置の実施の形態の一構成例を示すブロック図

【図２】本発明の文書グループ化装置の処理の流れの一実施の形態を示すフロー図

【図３】ＨＴＭＬ形式の文書の一例を示す図

【図４】ＨＴＭＬ形式の文書の一例を示す図

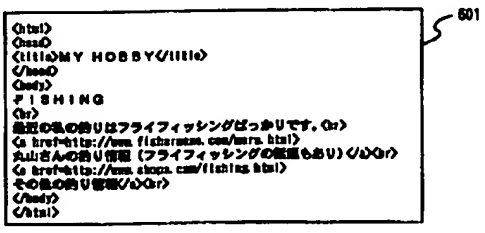
【図５】ＨＴＭＬ形式の文書の一例を示す図

【図６】本発明の文書グループ化装置の実施の形態の他の構成例を示すブロック図。

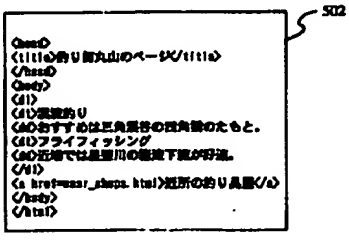
【符号の説明】

００１　文書グループ化装置
１１０　参照元文書格納手段
１２０　参照文書検出手段
１３０　文書取得手段
１４０　参照文書格納手段
１５０　文書グループ格納手段
１６０　関連性評価手段
５０１、５０２、５０３　ＨＴＭＬ形式の文書の例
９０１　文書群

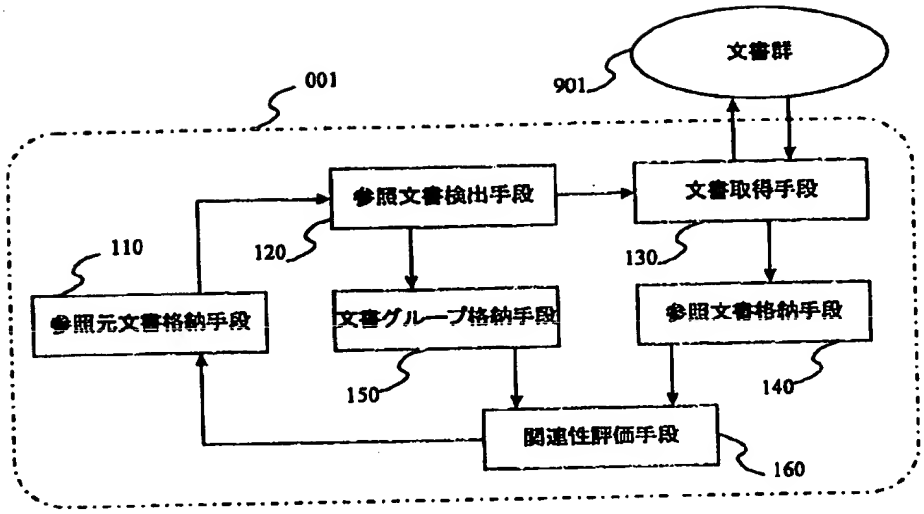
【図３】



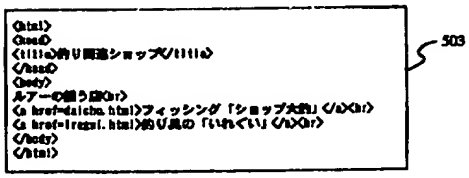
【図４】



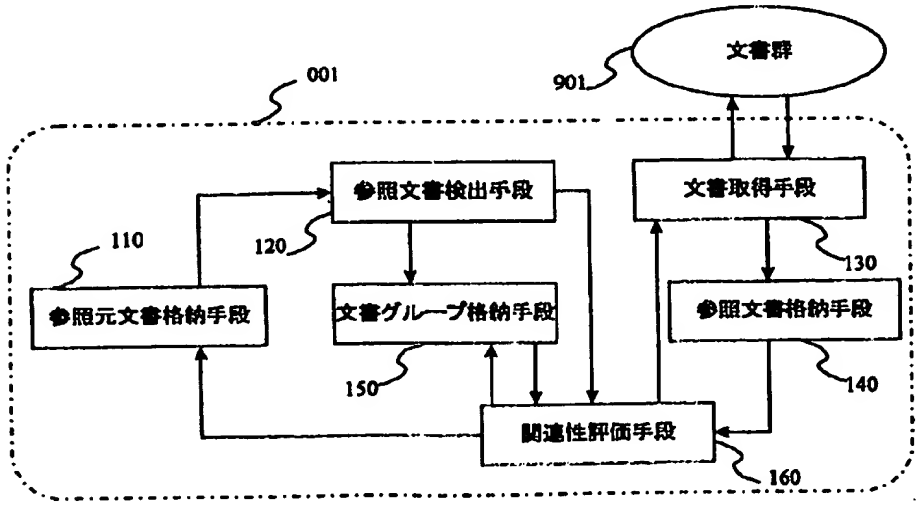
【図１】


文書群
901
001
参照文書検出手段
文書取得手段
110
120
130
参照元文書格納手段
文書グループ格納手段
参照文書格納手段
150
140
関連性評価手段
160

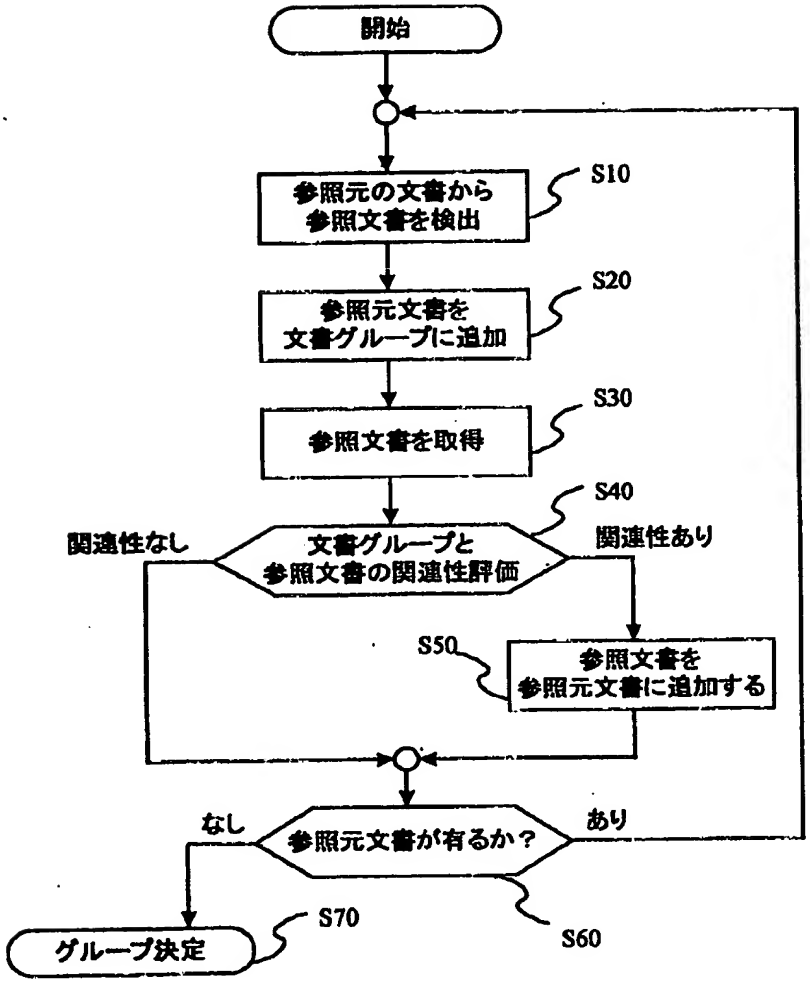
【図５】


503

【図６】


文書群
901
001
参照文書検出手段
文書取得手段
110
120
130
参照元文書格納手段
文書グループ格納手段
参照文書格納手段
150
140
関連性評価手段
160

【図２】


開始
参照元の文書から
参照文書を検出
S10
参照元文書を
文書グループに追加
S20
参照文書を取得
S30
文書グループと
参照文書の関連性評価
S40
関連性なし
関連性あり
S50
参照文書を
参照元文書に追加する
参照元文書が有るか？
なし
あり
S60
グループ決定
S70